Exploratory Analysis on Pixelwise Image Segmentation Metrics with an Application in Proximal Sensing

Paul Melki, Lionel Bombrun, Estelle Millet, Boubacar Diallo, Hakim Elchaoui Elghor, Jean-Pierre da Costa

remote sensing

MDPI

Article

Exploratory Analysis on Pixelwise Image Segmentation Metrics with an Application in Proximal Sensing

Paul Melki [1,2,*], Lionel Bombrun [1,3], Estelle Millet [2], Boubacar Diallo [2], Hakim ElChaoui ElGhor [2] and Jean-Pierre Da Costa [1,3]

[1] CNRS, IMS, UMR 5218, University of Bordeaux, F-33405 Talence, France; lionel.bombrun@ims-bordeaux.fr (L.B.); jean-pierre.dacosta@ims-bordeaux.fr (J.-P.D.C.)
[2] EXXACT Robotics, F-51200 Épernay, France; estelle.millet@exxact-robotics.com (E.M.); boubacar.diallo@exxact-robotics.com (B.D.); hakim.elchaoui@exxact-robotics.com (H.E.E.)
[3] Bordeaux Sciences Agro, F-33175 Gradignan, France
* Correspondence: paul.melki@exxact-robotics.com; Tel.: +33-635-151-620

Abstract: A considerable number of metrics can be used to evaluate the performance of machine learning algorithms. While much work is dedicated to the study and improvement of data quality and models' performance, much less research is focused on the study of these evaluation metrics, their intrinsic relationship, the interplay of the influence among the metrics, the models, the data, and the environments and conditions in which they are to be applied. While some works have been conducted on general machine learning tasks such as classification, fewer efforts have been dedicated to more complex problems such as object detection and image segmentation, in which the evaluation of performance can vary drastically depending on the objectives and domains of application. Working in an agricultural context, specifically on the problem of the automatic detection of plants in proximal sensing images, we studied twelve evaluation metrics that we used to evaluate three image segmentation models recently presented in the literature. After a unified presentation of these metrics, we carried out an exploratory analysis of their relationships using a correlation analysis, a clustering of variables, and two factorial analyses (namely principal component analysis and multiple factorial analysis). We distinguished three groups of highly linked metrics and, through visual inspection of the representative images of each group, identified the aspects of segmentation that each group evaluates. The aim of this exploratory analysis was to provide some clues to practitioners for understanding and choosing the metrics that are most relevant to their agricultural task.

Keywords: evaluation metrics; classification; image segmentation; proximal sensing; precision agriculture

Citation: Melki, P.; Bombrun, L.; Millet, E.; Diallo, B.; ElChaoui ElGhor, H.; Da Costa, J.-P. Exploratory Analysis on PixelwiseImage Segmentation Metrics with an Application in Proximal Sensing. *Remote Sens.* **2022**, *14*, 996. https://doi.org/10.3390/rs14040996

Academic Editors: Ganesh Bora and Dharmendra Saraswat

Received: 7 January 2022
Accepted: 15 February 2022
Published: 18 February 2022

1. Introduction

The performance evaluation of machine learning (ML) models is, from an applicative perspective, perhaps the most crucial step in the predictive pipeline as it is often framed as a decisional step [1]. Based on the computed evaluation metrics, the practitioner is called to decide whether a model is well-performing or not. Such a decision is often based on the "trust" attributed to the metrics computed and founded on two implicit assumptions: that the metrics are reflective of the true performance of the model on the test examples and that the metrics are reflecting the aspects of performance that are relevant to the application at hand. These two assumptions rely on an understanding of the theoretical and behavioural aspects of the evaluation metrics used, whose lack of clarity often increases with the complexity of the ML task (for example, classical binary classification metrics such as *precision* and *recall* are much easier to understand than their multiclass counterparts). This leads practitioners to adopt less elaborate evaluation metrics, whose numeric values are easier to interpret, report, and explain [2]. However, in complex tasks, the usage of simple metrics entails a simplification of the problem, leading to a loss of information about

the true performance of the model: the metrics may capture an aspect of performance that is not relevant to the application or are indeed evaluating a simplified version of the model. This can be quite problematic in real-world applications, where metrics offering a biased high positive evaluation of a model can lead practitioners to apply such a model under real-world conditions, only to find that it performs poorly, with a potential high risk [2–5].

While other topics are actively researched in the ML literature, such as the improvement of data quality [6–8] or the development of better-performing models [9–11], the metrics used to evaluate these predictive pipelines have taken a relatively minimal place in this field. Caruana and Niculescu-Mizil [12] provided one of the earliest comprehensive works on the topic, presenting nine performance metrics for binary classification, which they divided into three groups: threshold metrics, ordering/rank metrics, and probability metrics. The authors conducted an empirical analysis on these metrics using seven learning algorithms on a number of general-purpose classification datasets and then analysed the results using multidimensional scaling and correlation analysis. They also proposed a new evaluation metric named SAR, which combines the Squared error, Accuracy, and the Receiver Operator Characteristic (ROC). Building on this work, Alaiz-Rodriguez, Japkowicz, and Tischer [13] proposed that "classifier evaluation should be done on an exploratory basis", an approach we also take in the current work. Unlike the present research, the authors proposed an approach to visualise the performance of multiple models with the aim of comparing them, rather than aiming at understanding the relationships among these metrics. Seliya, Khoshgoftaar, and Van Hulse [14] applied this comparative study on a collection of 22 classification metrics applied on C4.5 decision tree classifiers trained with 35 datasets. They studied the relationships among these metrics using Common Factor Analysis (CFA) and identified four distinct groups of metrics. The authors concluded by suggesting that metrics that are highly related should not be used together as they do not add unique knowledge about the performance of a classifier. However, they did not provide a developed interpretation of why certain metrics are highly interrelated, while others are not. These surveyed works all studied the evaluation of classification models, with no particular context.

The application of ML models ranging from classical classification and regression algorithms, to state-of-the-art deep neural networks has played an important role in the advancement of remote and proximal sensing technologies, contributing to a wide range of domains. One such important applicative field is precision agriculture [11,15,16]. Indeed, advancements in ML techniques applied to computer vision have allowed the adoption of these proximal sensing technologies for the detection, identification, and measurement of plants and harvest [17]. An important task in this context is that of image segmentation defined as "the process of partitioning the image into semantically interpretable regions" [18], which permits the automatic detection and delimitation of plants in visual scenes. At its most basic level, image segmentation can be seen as a classification problem whereby the model's goal is to assign to each pixel in the image a specific class. Unlike other classification models, segmentation should not be seen as *exclusively* a classification problem, since the pixels classified are semantically constructed to form areas representing objects in the real world [19]. As such, even though classical classification metrics can be used to evaluate the performance of these models, it is important for practitioners to understand how these evaluations translate the semantic quality of the segmentation. In other words, it is important to understand how pixelwise classification quality impacts the visual result of segmentation, which may be considered *good* or *bad* depending on the needs. Scarce research works have studied evaluation metrics in the particular context of image segmentation, and none, to the authors' knowledge, have focused on the particular case of proximal sensing. Perhaps the most developed work on this topic is the paper by Taha and Hanbury [20], which presented and studied multiple evaluation metrics for 3D medical image segmentation. The authors defined the metrics and provided efficient algorithms to compute them. A correlation analysis was applied on the evaluation metrics computed in an empirical setup in order to identify the relationships among the metrics.

Most importantly, the authors provided interpretations of the obtained relationships and guidelines to help users choose the set of metrics that is most relevant to their application.

This work proposes a statistical methodology for analysing the relationships among pixel-level classification metrics in the context of image segmentation. It also provides concrete interpretations of these relationships in light of the segmentation masks produced by the models. The methodology is directly illustrated for the problem of segmenting plants from soil background in images obtained by proximal sensing in the context of automated weeding. The main objectives of the work can be summarised as follows:

- Provide a unified presentation of multiple pixel-level evaluation metrics for image segmentation;
- Define and develop a valid data mining exploratory approach that is both statistically valid and easily interpretable for the study of the relationships among these metrics;
- Provide a typological interpretation of the identified groups of metrics based on the observable segmentation results, with the aim of helping practitioners identify the useful metrics for their application.

The paper is organised as follows: In Section 2, we present the tools on which the analysis was conducted, namely: the data, the segmentation models we used for experimentation, and detailed definitions of the evaluation metrics, formulating them in terms of the four overlapping cardinalities of the confusion matrix. In Section 2.4, we detail the exploratory methodology applied, consisting of a correlation analysis, clustering of variables, Principal Component Analysis (PCA), and Multiple Factor Analysis (MFA). In Section 3, we present the obtained results of the first three methods applied on one chosen model. Individuals (images) were projected on the principal planes obtained in PCA, and their visual aspects were studied through a comparison of the Ground Truth (GT) masks and the constructed detection masks. We then show the results of the MFA applied on the metrics of multiple segmentation models with the aim of studying whether the metrics are "model-agnostic" or not. Finally, in Section 4, we comment on the obtained results and open the door to future work on the subject.

2. Materials and Methods

2.1. Experimental Dataset of Plant Images

The 153 images used to conduct the experiments were cut out of the original images of size 2940×1960, which were manually captured using a camera fixed on a monopod held perpendicularly to the ground under real-world conditions in outdoor fields. The images were gathered in multiple locations across France, presenting a variety of cultures under multiple conditions of luminosity. All the images were manually annotated by professional agronomist annotators, producing binary annotation masks, which served as GT masks in our evaluation of the segmentation models. From the originally obtained images, smaller crops of size 500×500 were manually produced in the lab, on regions of the image that contained some plants. We thus made sure that every image used in our experimentation contained some surface of plant in it, although the spatial distribution of these green regions and their sizes presented a rich variability among the images. The dataset was then randomly divided into a training set of 16 images and a test set of 137 images. Figure 1 shows a chosen sample of 4 images and their annotation masks, representing the variety of conditions in our database.

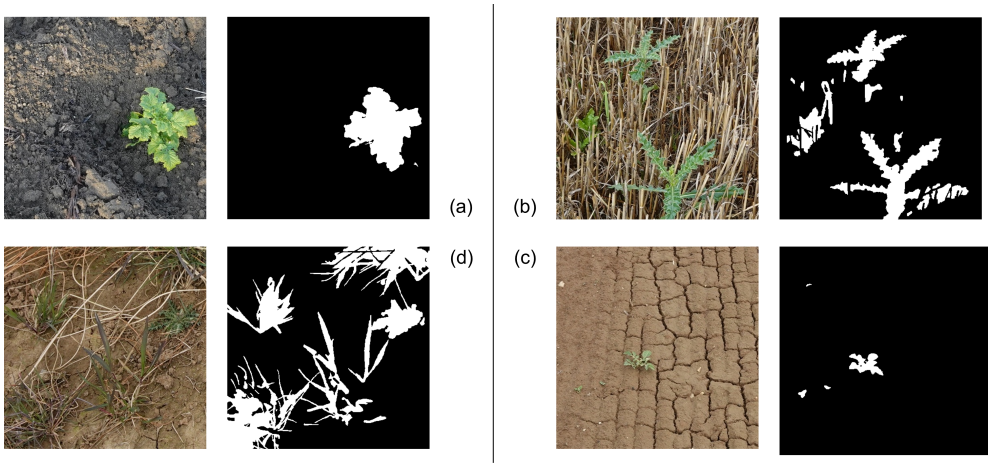

Figure 1. Random sample of 4 images and their annotation masks, showing the variety of conditions in the database. Clockwise starting from top-left: Image (**a**) shows an image with a big, clearly green plant, on a bar granular soil. Image (**b**) presents another variety of plant surrounded by dry grass, leading to partial occlusion of the plants' leaves, as is clearly apparent on the segmentation mask. Image (**c**) shows sparsely distributed small weeds on a crackled dry soil. Image (**d**) presents multiple weeds and broad-leaf plants on a bare residual soil.

2.2. Three Segmentation Models for Experimentation

The segmentation models we employed to study the image segmentation evaluation metrics were specific types of models that formulate the image segmentation problem as a pixel-level classification problem. That is, they aim to classify each pixel in an image as either a *plant* or a *background* pixel. From this low-level classification, we can reconstruct a segmentation mask, which we call the *detection mask*. A considerable number of models have been proposed in the literature to deal with the problem of pixel classification [21,22]. However, as the study and development of models are not the purpose of the present work, we experimented with only three models that have been shown in the literature to provide good results in agricultural contexts:

2.2.1. Decision Tree Segmentation Model

This model, adapted from [23], proposes to create a dataset of colour features extracted from the images, before training a classifier on them. The full image is traversed pixel by pixel, and the features are extracted. The colour features obtained are transformations of the *RGB* colours of each pixel into 6 different ordinal 3D colour spaces (*RGB*, *YCbCr*, *HSL*, *HSV*, *CIELab*, and *CIELuv*), from which we obtained 18 features for each observation (pixel). As some of the features that may be redundant or irrelevant can reduce the performance of the models, the original authors implemented a wrapper [24], which iteratively trains and tests the model on different combinations of features and keeps the subset of features that provides the best performance in terms of overall accuracy on an internal "test" set partitioned out of the training set. This led to retaining 9 features out of the original 18, which were the ones used in the construction of our model. For further details, the interested reader is referred to [23,24].

The classifier was an ordinary decision tree based on the well-known Classification and Regression Trees (CART) algorithm [25], which uses the *Gini Index* as its splitting function, or "measure of impurity".

2.2.2. Support Vector Segmentation Machine

Developed in [26], this approach proposes using *support vector machines* [27] for the problem of classifying pixels into *plant* and *background*. The authors proposed to take into consideration the neighbourhood pixel information in the *CIELuv* colour space to train an SVM classifier. The method can be summarised in this two-step procedure:

1. Transform the pixels from the original *RGB* colour space to the *CIELuv* colour space;
2. Train the SVM on the extracted features.

The first step of the procedure is of extreme importance since *CIELuv* is a perceptually uniform colour space, which means that in this space, Euclidean distances directly measure perceptual colour differences. From a geometric perspective, learning an SVM binary classifier is equivalent to finding the hyperplane in the feature space that best separates the two classes: as the distances between the points become clearer and easier to estimate, the learning process improves.

2.2.3. Colour Index of Vegetation Extraction

Similar to the previous models, the Colour Index of Vegetation Extraction (CIVE) method [28] aims at classifying pixels into either *plant* or *background* using a threshold separator. Unlike the previous models, however, this threshold is not "learned" from the extracted features, but is rather statically applied on one feature computed as a linear combination of the *RGB* intensities of each pixel:

$$Z = 0.441 \times R - 0.811 \times G + 0.385 \times B + 18.8 \tag{1}$$

A fixed threshold is applied on each pixel after computing its Z value, whereby pixels presenting a Z value greater than the threshold are classified as *plant* and the others as *background*.

2.3. Metrics

2.3.1. Overlap Cardinalities

As mentioned, the segmentation task can be formulated as a pixel classification problem, whereby the segmentation model is trying to learn, or define, a function M that maps each pixel p in an image to its correct class, either *plant* or *background*. From this low-level classification, we can spatially reconstruct a detection mask, which will be the segmentation produced by the model. Let $p = (r, c)$ be the pixel at row r and column c in a given image; the function M defining the colour intensity, and as a consequence, the class of each pixel can be defined as:

$$M(p) = \begin{cases} 0 & \text{if pixel } p \text{ belongs to a } \textit{background} \text{ object,} \\ 1 & \text{if pixel } p \text{ belongs to a } \textit{plant} \text{ object.} \end{cases} \tag{2}$$

Here, we can make the distinction between $M(p)$, the true class of pixel p, and $\hat{M}(p)$, the class predicted for this pixel by a given segmentation model. To simplify the notation in the rest of the work, we also define the GT mask containing the true labels as the $H \times W$ matrix $M_{H \times W}$, whereas the detection mask is defined as $\hat{M}_{H \times W}$.

Based on these definitions, we can derive the following confusion matrix of the pixelwise classification problem:

		Prediction		
		$\hat{M}(p) = 1$	$\hat{M}(p) = 0$	*Total*
GT	$M(p) = 1$	*TP*	*FN*	P
	$M(p) = 0$	*FP*	*TN*	N
	Total	$\hat{P}$	$\hat{N}$	n

Following the approach adopted in [20], we will refer to *TP* (True Positives), *FP* (False Positives), *TN* (True Negatives), and *FN* (False Negatives) as the *overlap cardinalities* and define the 12 evaluation metrics proposed in this work as functions of these quantities and their respective totals: P, which is the total of truly positive pixels, $\hat{P}$ which is the total number of pixels predicted as positive, and N and $\hat{N}$ their negative counterparts, or as functions

of the masks M and $\hat{M}$ when such a formulation is simpler. This approach simplifies the understanding of these metrics and their implementation, while demonstrating the link between low-level classification and higher abstraction segmentation.

2.3.2. Definition of the Metrics

We propose to study 12 evaluation metrics, some of which are well known and widely used in both the academic literature and industrial applications, while others are less reputed, probably due to the difficulty of their formulation and interpretation. We note all these metrics generally take values in $[0, 1]$ with the exception of the *Hausdorff distance*, which resides in $\mathbb{R}^+$:

1. *Positive predictive value: Precision* (PRC)
   This metric is the proportion of pixels actually labelled as plant among all the pixels predicted as plant. It allows us to see the precision of the model in predicting plant pixels.
   $$PRC = \frac{TP}{TP + FP}; \tag{3}$$

2. *Negative Predictive Value* (NPV)
   This metric shows the proportion of pixels correctly classified as background among all the pixels classified as background. This allows us to see the precision of the algorithm on the background pixels.
   $$NPV = \frac{TN}{TN + FN}; \tag{4}$$

3. *Recall* (RCL)
   Known also as the *true positive rate*, this is the proportion of pixels correctly classified as *plant* among all the pixels that are in reality *plant* pixels.
   $$RCL = \frac{TP}{TP + FN}; \tag{5}$$

4. F_β*-Score* (F1S)
   This metric combines *precision* and *recall* into one evaluation metric parameterised by β, which specifies how much more we are interested in the *recall* than in the *precision* [29]. Formally, the F_β-score is defined as:
   $$F_\beta = (1 + \beta^2) \times \frac{PRC \times RCL}{\beta^2 \times PRC + RCL} \tag{6}$$
   The most common version used is the F_1-score, which assigns the same importance to *precision* and *recall* with $\beta = 1$:
   $$F_1 = 2 \times \frac{PRC \times RCL}{PRC + RCL}; \tag{7}$$

5. *Accuracy* (ACC)
   Probably one of the most widely used evaluation metrics for classification problems, this quantity shows the proportion of correct decisions among all decisions made, that is the proportion of pixels correctly classified, among all the pixels in the image:
   $$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{n}; \tag{8}$$

6. *Balanced Accuracy* (BAC)
Introduced in [4] with the aim of solving the problem faced by the *accuracy*, the *balanced accuracy* can be defined as:

$$BAC = c \times \left(\frac{TP}{TP+FN}\right) + (1-c) \times \left(\frac{TN}{TN+FP}\right) \tag{9}$$

where $c \in [0,1]$. In this case, we define c as the cost associated with the misclassification of a *positive* example, which gives us the freedom to set the penalisation in the case of misclassification on our class of interest. This may prove to be extremely important in certain use cases where there is a considerable loss associated with the misclassification of positive cases. In this work, we take $c = \frac{1}{2}$ in order to give equal weights to the two classes;
7. *Intersection Over Union (Jaccard Index)* (IOU)
Widely used in the computer vision literature on a variety of vision tasks including object detection and image segmentation, the *Intersection over Union* (IoU)—also known as the *Jaccard Index*—is perhaps the most famous spatial overlap metric in usage. For two finite sets A and B, it can be defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

where $|\cdot|$ is the set cardinality operator. Taking the two sets of interest to be our *ground truth* and *detection* masks, we can measure their similarity as:

$$J(M,\hat{M}) = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|} \tag{11}$$

It was shown in [20] that in the case of binary classification, where the two sets studied are the sets of elements being classified (in our case, the pixels), the Jaccard Index can be easily reformulated in terms of the overlap cardinalities as:

$$J(M,\hat{M}) = IOU(M,\hat{M}) = \frac{TP}{TP+FP+FN}; \tag{12}$$

8. *Global Consistency Error* (GCE)
The GCE [30] is a global measure of the segmentation error between two segmentation masks, based on the aggregation of local consistency errors measured at each pixel. Being a measure of error, it is a metric to be minimised. Defining $R(M,p)$ as the set of all pixels that belong to the same class as the pixel p in the segmentation M, then we can define the segmentation error between two segmentations M and $\hat{M}$ at p as:

$$E(M,\hat{M},p) = \frac{|R(M,p) \setminus R(\hat{M},p)|}{|R(M,p)|} \tag{13}$$

where $\setminus$ is the set difference operator. As such, the GCE is defined as the error averaged over all the pixels and given by:

$$GCE(M,\hat{M}) = \frac{1}{n} \min\left\{\sum_p E(M,\hat{M},p), \sum_p E(\hat{M},M,p)\right\} \tag{14}$$

It can be shown that the GCE can be re-written in terms of the overlap cardinalities as [20,30]:

$$GCE = \frac{1}{n} \min\left\{\frac{FN(FN+2TP)}{TP+FN} + \frac{FP(FP+2TN)}{TN+FP}, \frac{FP(FP+2TP)}{TP+FP} + \frac{FN(FN+2TN)}{TN+FN}\right\}; \tag{15}$$

9. *Relative Vegetation Area* (RVA)
This metric, specific to the case of binary segmentation and proposed in [31], is a simple spatial overlap metric, which compares the area of vegetation detected in the segmentation mask to the true detection area in the ground truth mask. It is defined as:

$$RVA = \begin{cases} 1 - \frac{P-\hat{P}}{P} = 1 - \frac{FN-FP}{TP+FN} & \text{if } \hat{P} < P \\ 1 - \frac{\hat{P}-P}{\hat{P}} = 1 - \frac{FP-FN}{TP+FP} & \text{if } P < \hat{P} \end{cases} \tag{16}$$

where $P = TP + FN$ is the vegetation area (number of plant pixels) in the ground truth mask and $\hat{P} = TP + FP$ is the vegetation area in the detection mask;

10. *Adjusted Mutual Information Index* (AMI)
The mutual information of two random variables X and Y is a quantification of the amount of information that X holds about Y. It measures the reduction in uncertainty about Y given knowledge of X. The *mutual information index* of two segmentation masks M and $\hat{M}$ can be understood as a measure of the amount of *true* information that the detection mask produced by the algorithm contains, in comparison to the information contained in the GT mask. It is based on the marginal entropies $H(M)$, $H(\hat{M})$ and the joint entropy $H(M, \hat{M})$, which are defined as:

$$H(M) = -\sum_{i=0}^{1} \pi(M^i) \log \pi(M^i) \tag{17a}$$

$$H(\hat{M}) = -\sum_{i=0}^{1} \pi(\hat{M}^i) \log \pi(\hat{M}^i) \tag{17b}$$

$$H(M, \hat{M}) = -\sum_{i=0}^{1} \sum_{j=0}^{1} \pi(M^i, \hat{M}^j) \log \pi(M^i, \hat{M}^j) \tag{17c}$$

where $\pi(M^i)$ is the probability of observing a pixel of class i in the mask M, which can be expressed in terms of the four overlap cardinalities as:

$$\pi(M^1) = \frac{P}{n} \qquad \pi(M^0) = \frac{N}{n} \qquad \pi(\hat{M}^1) = \frac{\hat{P}}{n} \qquad \pi(\hat{M}^0) = \frac{\hat{N}}{n} \tag{18}$$

and the joint probabilities $\pi(M^i, \hat{M}^j)$ as:

$$\begin{aligned} \pi(M^1, \hat{M}^1) &= \frac{TP}{n} & \pi(M^1, \hat{M}^0) &= \frac{FN}{n} \\ \pi(M^0, \hat{M}^1) &= \frac{FP}{n} & \pi(M^0, \hat{M}^0) &= \frac{TN}{n} \end{aligned} \tag{19}$$

and finally, the *Mutual Information Index (MI)* is defined as:

$$MI(M, \hat{M}) = H(M) + H(\hat{M}) - H(M, \hat{M}) \tag{20}$$

Computationally, we compute the *Adjusted Mutual Information Index* (AMI) as developed in the `scikit-learn` Python library [32], which is the version of the MI adjusted for chance [33].

11. *Cohen's Kappa Coefficient* (KAP)
On the probabilistic side, we implemented *Cohen's Kappa coefficient*, which is a measure of agreement between two samples. As an advantage over other measures with the exception of the AMI, the Kappa takes into account the agreement caused by chance, thus making it a fairer measure of model performance. It can be defined as:

$$KAP = \frac{P_a - P_c}{1 - P_c} \tag{21}$$

where "P_a is the agreement between the two samples" (simply, ACC), and "P_c is the hypothetical probability of chance agreement" [20]. We can re-write the Kappa using frequencies, which in turn can be expressed using the four overlap cardinalities, in order to facilitate our computations:

$$KAP = \frac{f_a - f_c}{n - f_c} \tag{22a}$$

$$\begin{aligned} f_a &= TP + TN \\ f_c &= \frac{(TN + FN)(TN + FP) + (FP + TP)(FN + TP)}{n}; \end{aligned} \tag{22b}$$

12. *Hausdorff Distance* (HDD)
Originally, this metric was formulated to measure the dissimilarity between two subsets of a metric space [34]. Informally, it allows measuring the degree of mismatch between two finite sets by measuring the distance from the point of Set 1 that is farthest from any point of Set 2, and vice versa. For example, if the distance between the two sets is k, then every point in one set must be within a distance k from every point in the other set [35]. This metric has been previously introduced and widely applied in the computer vision literature, especially in the field of medical imagery [20,35]. Formally, for two finite sets M and $\hat{M}$, we define the Hausdorff distance as:

$$d_H(M, \hat{M}) = \max\left\{ \sup_{m \in M} d(m, \hat{M}), \sup_{\hat{m} \in \hat{M}} d(M, \hat{m}) \right\} \tag{23}$$

where $d(x, Y) = \inf_{y \in Y} d(x, y)$ and d is some distance metric, which is usually, and in our case, the Euclidean distance.
The closer the two objects M and $\hat{M}$ are to each other in the Hausdorff distance, the more similar they are in shape. That is, two segmentation masks that have a low Hausdorff distance are masks in which the pixels are similarly distributed spatially over the grid. A more detailed formulation about the computation of the Hausdorff distance for comparing images can be found in [35].

2.4. Methodology for the Exploration of Metrics' Relationships

The trained models were used to compute predictions on the test images from which the detection masks were constructed. The 12 evaluation metrics were computed three times (one for each model) for each test image through the comparison of the GT and detection masks, and an exploratory study was conducted on these metrics.

The exploratory approach first focused on studying the variability of the metrics of the images for each model separately, based on correlation analysis, principal component analysis, and clustering of variables, which require the data to satisfy certain conditions. Indeed, it is important to account for the important discrepancies among the distributions of the raw metrics, whereby some metrics present highly skewed distributions, while others follow a more symmetric distribution, as can be seen in Figure 2. It is known that correlation-based methods can be highly influenced by skewed distributions, which will bias the results in the direction of their skew, and by outlying observations [36]. To account for this, we transformed our dataset of raw metrics into a dataset of rankings by metric, wherein for each observation (image), we have the ranking of this observation in the full dataset given by each metric (column). In this way, we would have transposed our data into a homogeneous space in which the "decisions" made by the metrics can be compared and analysed. This transformation additionally allowed us to account for the fact that some of the metrics are distances or errors (to be minimised), while others are similarities (to be maximised), which would have made the interpretation of the results slightly more difficult. We note that this method was also used in [20].

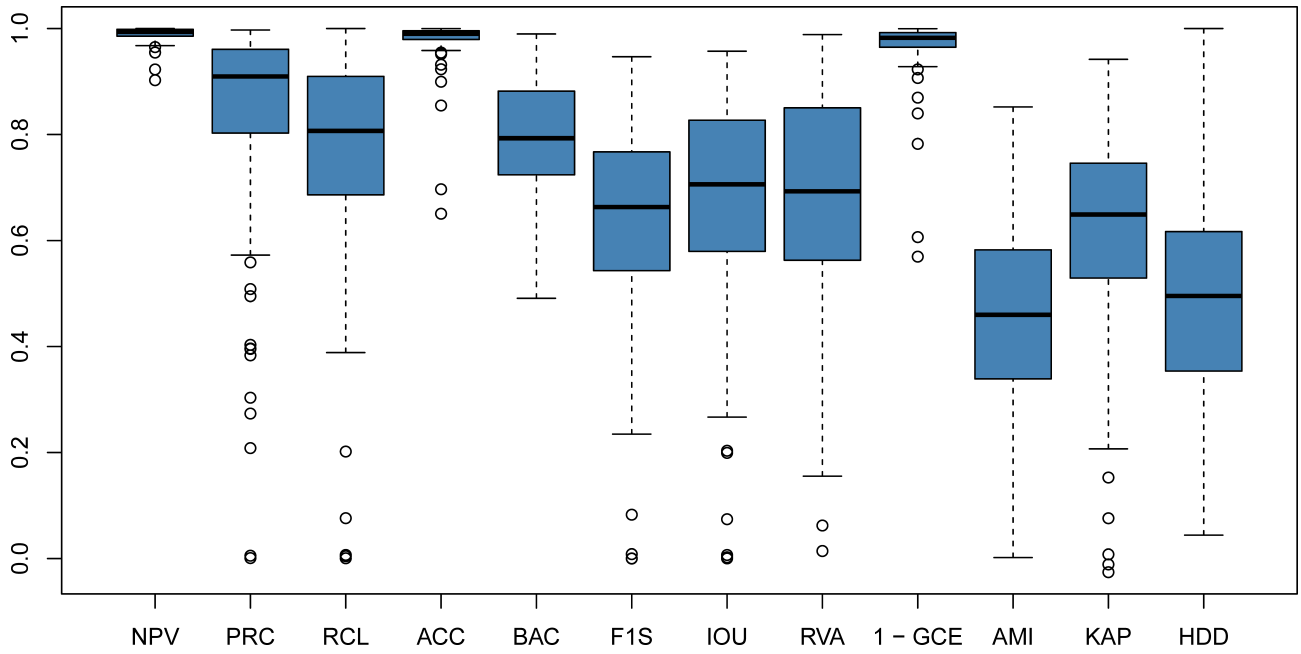

Figure 2. Boxplots of the raw metrics' values showing the discrepancies in their distributions. The HDD, which takes values in $\mathbb{R}^+$, is transformed as $HDD/\max(HDD)$.

The exploratory approaches taken to study the metrics were the following:

1. *Correlation analysis*
   The first analysis conducted was the study of the correlation matrix computed based on *Spearman's rank correlation*. Spearman's correlation ρ is a well-known nonparametric measure of rank correlation, showing how well the relationship between two variables can be described using a monotonic function, regardless of whether this relationship is linear or not [37]. It is clear that Spearman's correlation is simply the Pearson correlation applied to the rank variables. In our case, two metrics that have a high Spearman correlation are thus consistently ranking the same image similarly, as a good-quality image with a high ranking in the dataset, or otherwise. As such, it can be seen as a measure of "agreement" between the metrics;
2. *Clustering of variables*
   Another elaborate method for the exploration of relationships among variables is *clustering of variables* [38]. This method provides results that are clear to understand, interpret, visualise, and report. Given a set $\{x_1, ..., x_p\}$ of quantitative variables, the aim of this method is to find the partition P_K of these variables into K clusters $\{C_1, ..., C_K\}$ in such a way so as to maximise the "homogeneity" of the partition, which indeed amounts to maximising the correlations among the variables of the same cluster. The clustering method used was hierarchical clustering, aiming at building a set of p nested partitions of variables following the algorithm detailed in [38]. This method was implemented in `R` using the package `ClustOfVar` [38];
3. *Principal Component Analysis (PCA)*
   To further confirm our results and to provide a clearer visualisation of the results, we conducted a *Principal Component Analysis* (PCA) [39,40] and plotted the obtained correlation circles. PCA aims at constructing new variables that are linear combinations of the original ones, under the constraints that all Principal Components (PCs) be pairwise orthogonal, with the first PC explaining the largest part of the dataset's inertia, that is having the greatest "explicative" power, the second PC having the second largest inertia, and so on. The correlation circle that shows the projection in 2D of the variables with respect to some chosen PCs is a well-known and easily interpretable result of PCA, which allows us to visually identify groups of highly

correlated variables among each other, and with particular PCs. This analysis was conducted in `R` using the package `FactoMineR` [41];

4. *Multiple factor analysis*
   In Sections 3.1.1 and 3.1.2, the results of the three exploratory methods are shown only for the decision tree segmentation model, for conciseness. However, in order to verify that the results we obtained were not model dependent, we applied a Multiple Factor Analysis (MFA) [42,43] on the rankings produced by the three models. The result of this analysis would show us whether the relationships among the metrics differ largely from one model to another, or otherwise. From a simple perspective, the MFA can be understood as a PCA applied on the principal components obtained by the separate PCAs for each model, although the method is more elaborate. The most valuable result produced by MFA for this study was the representation of the *partial axes*. It consists of visualising the projection of the first and second dimensions obtained in each separate group PCA, on the principal plane of the MFA, formed by the first and second principal vectors of the MFA [42]. Such a method allows visualising how well correlated each of the most important dimensions of each group are with their respective global ones, in addition to showing their correlation among each other. The MFA was implemented in `R` using the package `FactoMineR` [41].

3. Results

3.1. Results of the Analysis on One Model

3.1.1. Correlation Analysis

The results of the first analysis using Spearman's correlation coefficient suggested three clear distinct groups of highly correlated metrics. These are the metrics that gave consistently similar rankings of quality across all test images. Figure 3 shows the correlation heat map obtained for the DTSM, along with the three groups of metrics identified. The three groups are hereby identified by their assigned colours: *green*, *orange*, and *blue*.

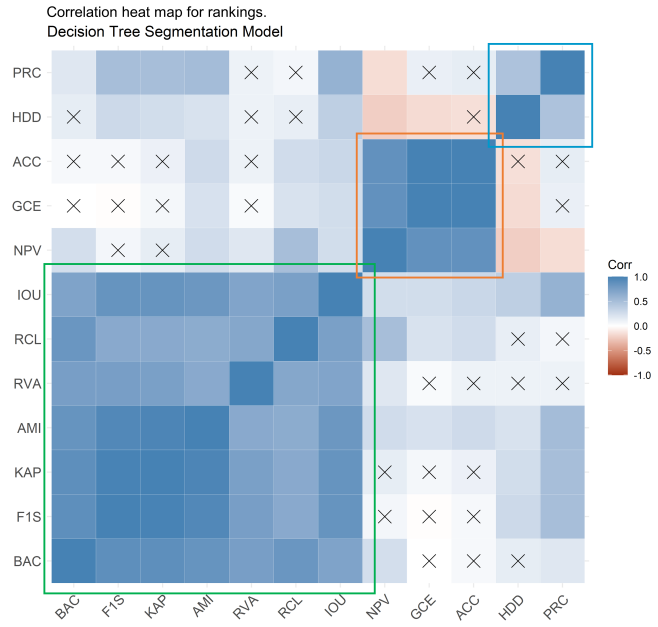

Figure 3. Spearman's correlation heat map showing the three groups of highly correlated metrics.

3.1.2. Clustering of Variables

Applying the clustering of variables allowed us to clearly identify distinct groups of metrics. Three groups of metrics were identified based on the results shown in Figure 4 where the optimal number of clusters can be seen to be three, in Panel (a), since it is the number of clusters after which the loss of homogeneity becomes quite considerable for

additional clusters. This is also confirmed in Panel (b), showing the value of the mean of the adjusted Rand index computed by comparing 40 bootstrap samples and the initial hierarchical partition. The maximum value appears to be at three clusters, signifying that this is the most stable level of partitioning [38].

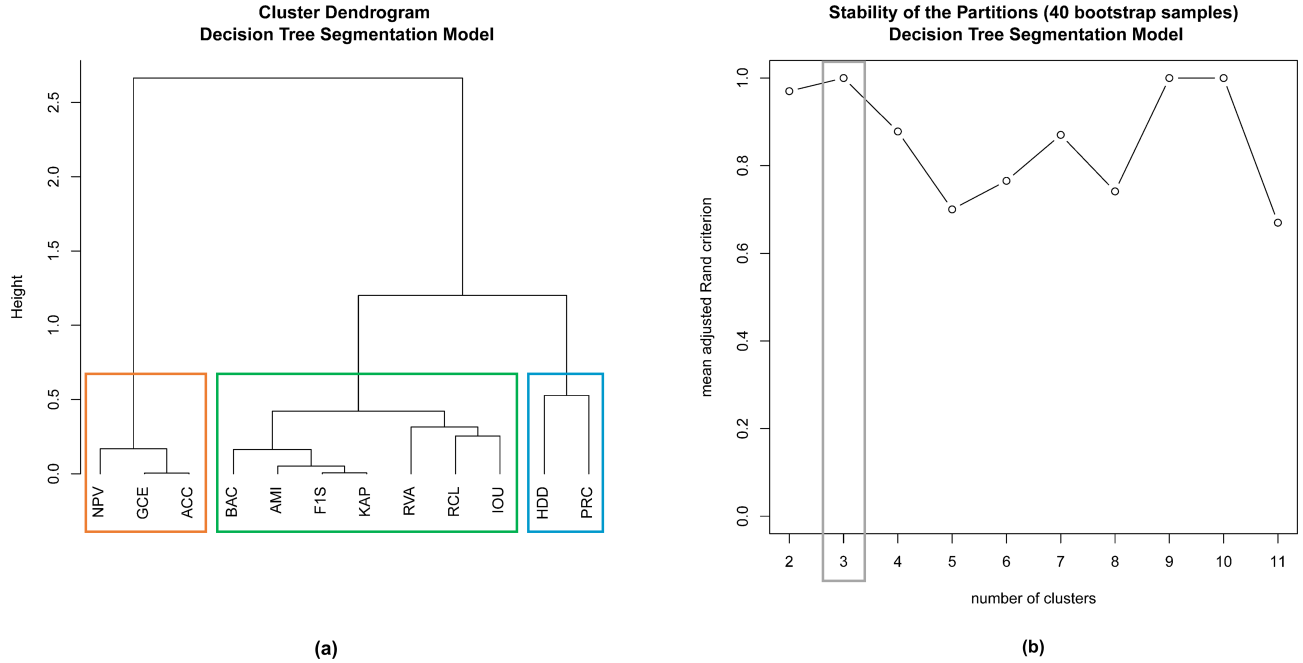

Figure 4. (**a**) Dendrogram resulting from hierarchical clustering. Height (*y*-axis) represents the loss in homogeneity. (**b**) Measure of stability over 40 bootstrap samples in comparison with initial clustering: we can see that the partitioning is most stable at 3 clusters.

As the presented results clearly showed, the twelve metrics studied can be separated into three distinct groups: the *green* group consisting of Balanced Accuracy (BAC), Recall (RCL), F1-Score (F1S), Cohen's Kappa (KAP), Intersection Over Union (IOU), Relative Vegetation Area (RVA), and Adjusted Mutual Information (AMI); the *orange* group consisting of Negative Predictive Value (NPV), Global Consistency Error (GCE), and Accuracy (ACC); and the *blue* group consisting of Precision (PRC) and the Hausdorff Distance (HDD). Even though the statistical methods proposed allow for the recognition of a pattern among the metrics, they do not explain precisely why certain metrics are in the same group, while others in a different group. The goal of the study was not only to analyse the structure of the evaluation metrics, but also to try to understand this in the specific context of the segmentation of images obtained by proximal sensing.

3.1.3. Principal Component Analysis

We focused on the correlation circle obtained by PCA, which shows the projection of each metric on a plane formed by two chosen principal vectors. Most practitioners focus on the plane formed by the first and second principal components, being the ones that explain the highest proportion of the data's inertia, thus showing the greatest amount of information. However, looking at different planes formed by less-important principal components in certain situations revealed more clearly some patterns that may not appear very distinctly by studying exclusively the principal plane formed by the first and second principal vectors. The original dimension of the dataset being twelve (for the twelve evaluation metrics), we applied PCA in order to reduce its dimension down to three, wherein the first three principal components retain around 88% of the original dataset's inertia. It is possible to identify in Figure 5, Panel (a), the three groups of metrics identified in the correlation analysis. Indeed, the metrics forming the *green* group are all clearly highly correlated with the first principal component, with low correlations with the second

principal component. The two other groups present a lower correlation with the first principal component, suggesting that these metrics are evaluating the images differently than the *green* group.

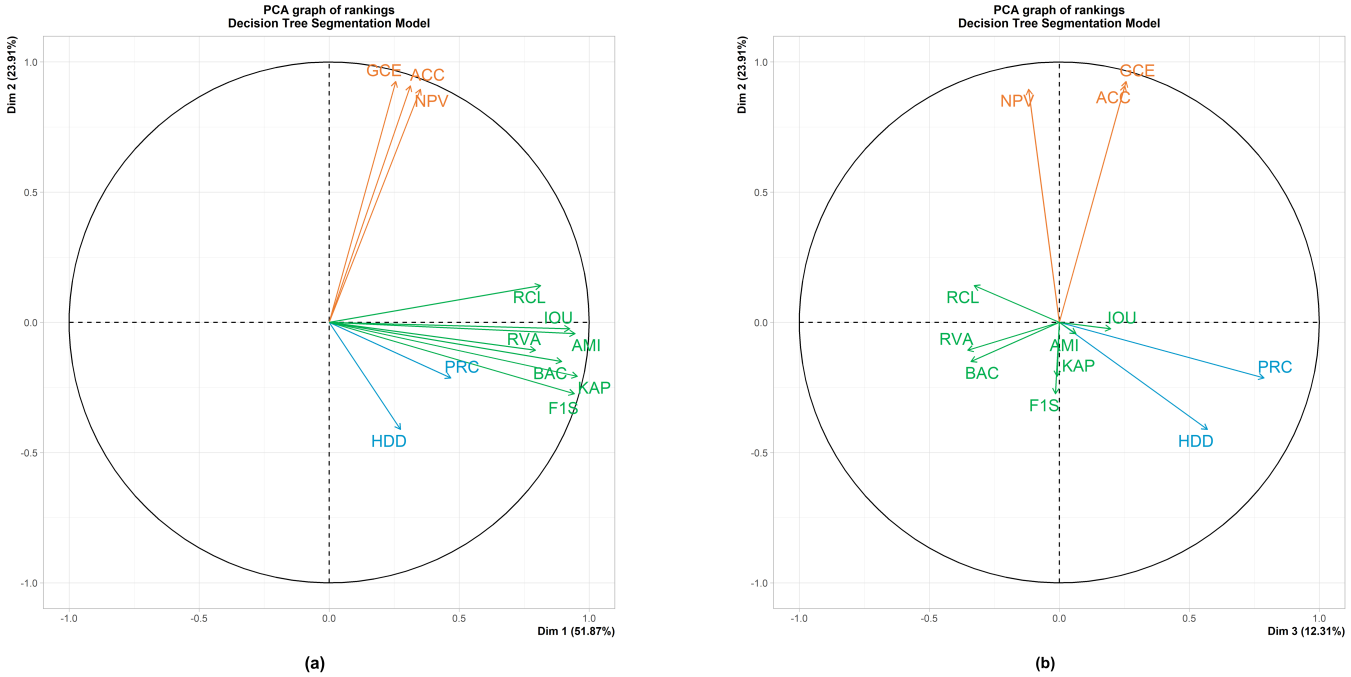

Figure 5. Projections of the metrics on the planes formed by: (**a**) 1st and 2nd PCs; (**b**) 3rd and 2nd PCs.

However, the *orange* and *blue* groups are clearly uncorrelated along the second PC. This divergence is more evident in Panel (b), where the *orange* group shows a very high correlation with the second PC, while the *blue* group has a high positive correlation with the third principal component. Indeed, it can be concluded that the three groups of metrics, *green*, *orange*, and *blue*, can be summarised or explained by the *first*, *second*, and *third* components, respectively, and are uncorrelated among each other.

3.2. Visual Inspection of the Segmentation Masks

A visual inspection of manually chosen images that are representative of each group of metrics provides valuable insight for understanding the obtained groups. For this, we looked at the projections of individuals along the principal plane formed by the first and second principal components obtained by the PCA in Section 3.1.3. As each principal component $s \in \{1, ..., S\}$ is a linear combination of the $K = 12$ original variables, each individual (image) residing in $\mathbb{R}^K$ can be projected into the lower space $\mathbb{R}^s$, which for the sake of simplifying visualisations was taken to be $\mathbb{R}^2$ [40,42,44]. As the metrics that originally resided in $\mathbb{R}^n$ were also projected in $\mathbb{R}^2$, the individuals overlapping or projected along the same direction of that of a group of metrics as visualised in Figure 5 were the images that were highly ranked by this group of metrics, while the individuals projected on the opposite side were the ones that were not well evaluated by this group of metrics.

Based on this, the rightmost images (in dark green in Figure 6) are supposed to be the ones that are highly evaluated by the *green* group of metrics, while those projected on the leftmost side (in light green) are the ones with the worst evaluations according to this group of metrics. Under the assumption that the metrics we used provide a correct evaluation of the quality of segmentation, the rightmost images would present detection masks that are very similar to the GT masks. Indeed, in Figures 6 and 7, showing some representative images, we can see that the two rightmost images, namely Images 88 and 137, are very well

segmented by the model. There is practically no difference between the GT mask and the detection mask.

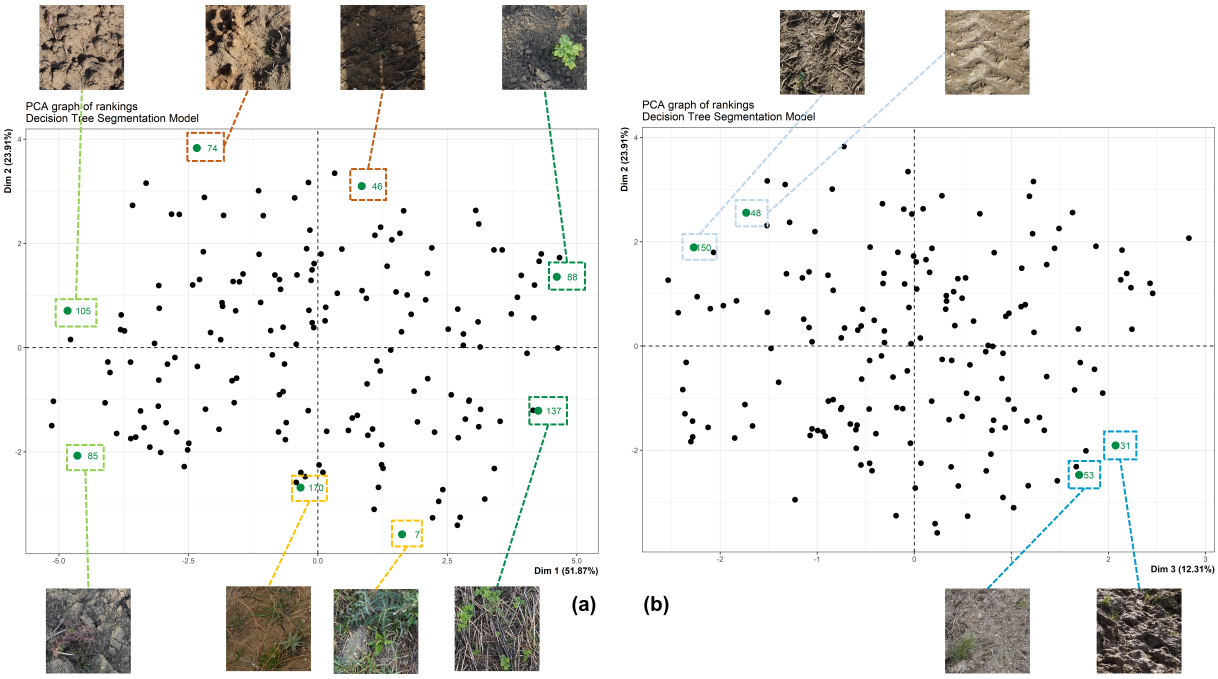

Figure 6. Representative images for each group of metrics and their locations in the projected point clouds in the principal planes formed by: (**a**) 1st and 2nd PCs; (**b**) 3rd and 2nd PCs.

These good segmentations can be contrasted with the leftmost images, the supposedly worst evaluated images by the green metrics, of which we exhibit Images 85 and 105. It appears clearly that these images were very badly segmented with a very high rate of error. The model, for these images, can hardly be said to have detected the plants of interest, since significantly large areas of the objects were not segmented. This suggests that, indeed, the group of *green* metrics provides correct and reliable evaluations of the quality of segmentation and discriminates well the good segmentations from the bad ones.

Another clear pattern that can be identified is the difference in the sizes of the objects in the images at the top of the point cloud projected on the first principal plane in comparison with those at the bottom side. The images at the top, whether they are well segmented (on the right, as 46), or badly segmented (on the left, as 74), all contain small plants (Figures 6 and 7). On the other hand, we can clearly see that on the bottom-side images, the plants take up a considerable area in the images. To understand this contrast in sizes, we have to reason in relation to the proportion of plant pixels in this context of binary classification and the group of metrics giving a high evaluation for one group or the other. Indeed, the size of the plants in the images can be simply understood as the proportion of plant pixels in the image, that is the positive cases in terms of binary classification. When the plants are very small, there is a drastic imbalance between the *plant* and *background* classes, while in images that present a higher plant surface, the two classes are better balanced. Accordingly, the images at the upper side of the point cloud are those that were highly evaluated by the *orange* group of metrics, which was highly positively correlated with the second dimension. It is well known and was shown in [4] that the accuracy does not provide reliable results when there is considerable imbalance between the classes. In particular, when both training and test sets present the same direction of imbalance—which is the case here, as the majority of images had more *background* than *plant*—the accuracy will yield an optimistic evaluation. It is also for this reason that the NPV gave a high evaluation, since it is a metric that emphasises TNs, which are much more likely to appear as the negative cases form the majority class.

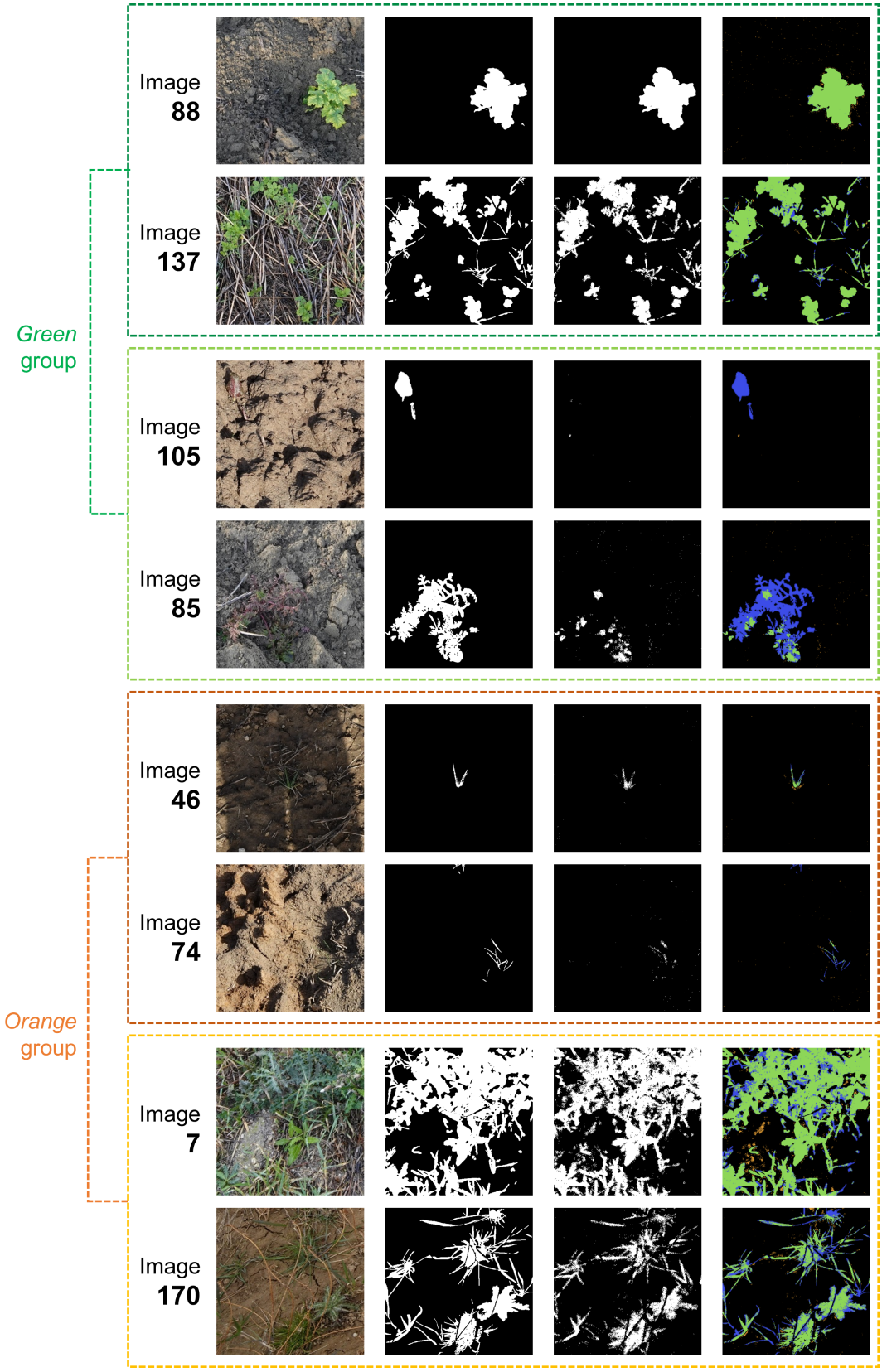

Figure 7. The representative images shown in Figure 6a with their GT, detection, and comparison masks showing the TP, FN, and FP predictions.

In the projection of the point cloud on the principal plane formed by the second and third principal vectors (Figure 6b), we can see that the images that were projected on the

upper left corner of the cloud—such as Images 48 and 150—were the images that presented a high number of FPs (Figure 8). These false detections were mostly of a very small size and highly dispersed around the image, in a random granular texture. These images happen to be those that were badly evaluated by the *blue* group of metrics. Indeed, as precision measures the proportion of TP predictions among all positive predictions, an image presenting a high number of FPs will reduce the precision, since a considerable number of these positive predictions are incorrect ones. Take the extreme situation of a detection mask that is completely white—that is, in which all pixels are predicted as *plant*: such a detection mask will have a perfect recall of 100% since all plant pixels have indeed been segmented, but a low precision due to the high number of FPs. These images were also badly evaluated by the Hausdorff distance because it is a spatial metric that finds for every pixel in one mask its closest pixel having the same class in the other mask. As such, all the highly granular small falsely predicted *plant* pixels, which are far away from the true object, will lead to a very fast explosion of the value of the computed distance, leading to a bad evaluation of such a mask. The images projected on the opposite side—namely, Images 31 and 53—were the ones in which every GT area was partially detected, but not fully. These images were highly evaluated by both the PRC and HDD due to the good proportion of TPs they identified and to the fact that the detected pixels were always *inside* the GT area. In the context of detection, the objects may be considered as detected, even though the segmentations proposed by the model were far from perfect.

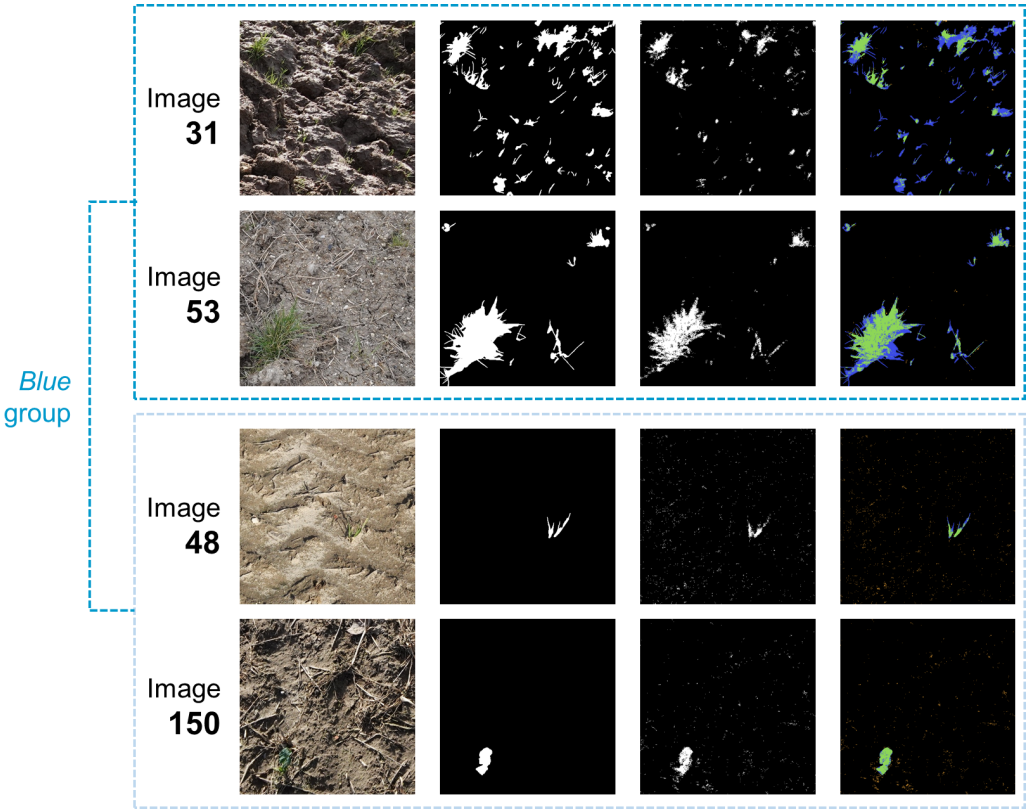

Figure 8. The representative images shown in Figure 6b with their GT, detection, and comparison masks showing the TP, FN, and FP predictions.

3.3. Multiple Factor Analysis

By looking at the plot of the partial axes shown in Figure 9, in which we see the projections of the first, second, and third PCA dimensions for each separate group (model), on the first, second, and third dimensions of the MFA, it is clear that the PCA dimensions were very closely projected for the three models. This suggests that the PCAs applied on each model separately were finding very similar structures in the data and were yielding

similar projections of the metrics on the principal plane. This supports the assertion that the analytical procedure applied on one model was indeed model independent and can be generalised to other classification models. These results also suggest that the segmentations obtained by the three models were not very different and that, in the given conditions, the three models generally produced similar results.

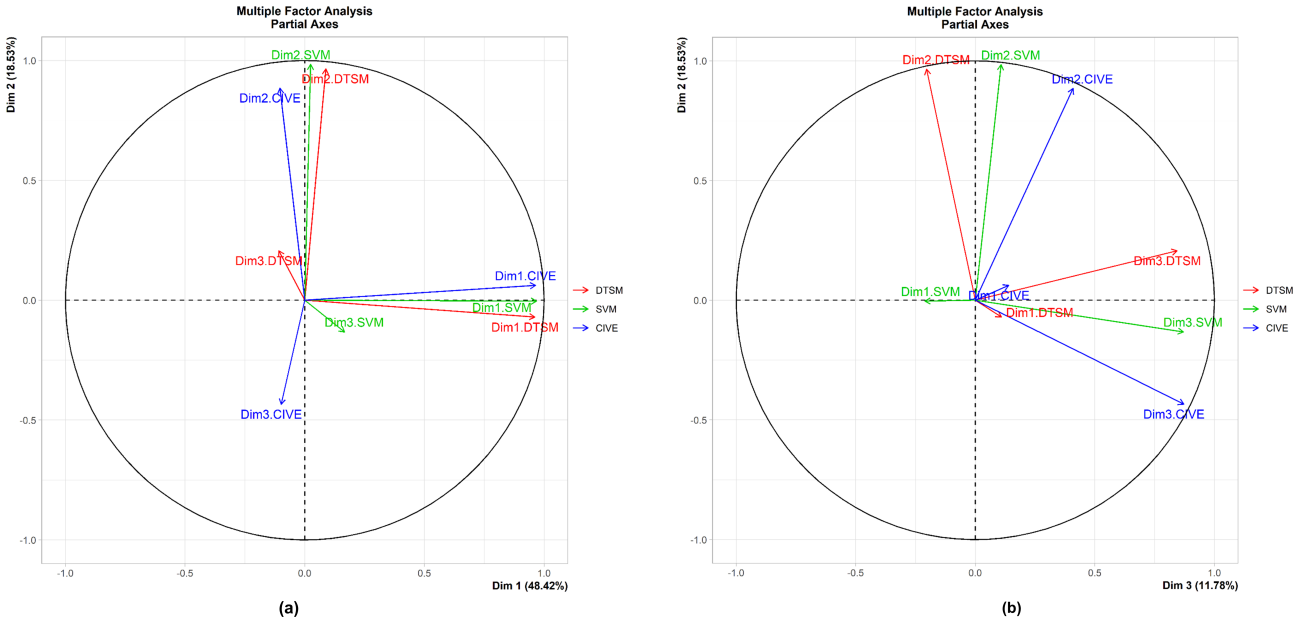

Figure 9. MFA partial axes' plot: (**a**) 1st and 2nd dimensions; (**b**) 3rd and 2nd dimensions.

4. Conclusions

In this paper, we presented and analysed 12 classification metrics in the context of plant image segmentation. After defining the metrics as functions of the overlap cardinalities, we studied the behaviour of these metrics by looking at the rankings of the test images computed based on these metrics. An exploratory analysis starting with a correlation analysis, followed by PCA and clustering of variables, led to the identification of three distinct groups of metrics. Namely, the first and bigger group consisted of balanced accuracy, IOU, F1-score, recall, adjusted mutual information and Cohen's Kappa. This group of metrics has proven to be well-performing regardless of the intrinsic structure of the image. In particular, their evaluation seemed independent of the imbalance between the two classes. As such, these metrics can be used for general-purpose segmentation tasks, where exact segmentation with well-delineated contours is not of the utmost importance. Practitioners need not compute all these metrics, but can choose one or more representative metrics from this group. It is always encouraged, however, not to rely on only one evaluation metric. Thus, practitioners can adopt certain metrics from this group, which measure the "goodness" differently: for example, IOU for overlap, Cohen's Kappa to account for randomness, and adjusted mutual information for an information-theoretic measurement. Even though these metrics provide conforming rankings of the images, their returned raw values may vary in distribution, with some showing highly positively skewed distributions, which make the distinction of quality between images harder to identify through a simple reading of their numeric value. As such, it is always encouraged to include some less-"optimistic" metric such as Cohen's Kappa in one's evaluation, for a more sober evaluation.

The second group of metrics consisted of the accuracy, negative predictive value, and global consistency error. These metrics have been found to be highly sensitive to class imbalance. Indeed, the accuracy, which is a widely used metric due to its ease of computation and interpretation, should only be used in cases where the classes are roughly balanced. Otherwise, it will tend to be biased towards the majority class. Similarly, the NPV, which can be understood as the "precision" on the negative class, is also highly

sensitive in cases where this class is the majority one. However, in situations where it is important to take into account the performance on the negative class, the NPV provides an interesting measure, given that a balance between the two classes is established. As for the GCE, this measure is known to be tolerant to refinements, meaning that there are two trivial segmentations that achieve zero error: one pixel per segment and one segment for the entire image, since the first one is a refinement of any other segmentation and any segmentation is a refinement of the second [30]. As such, in the cases where there is a clear imbalance between the classes and the model is biased in the direction of the majority class, this metric will tend to the zero error, since the masks will also tend to approach a mask that contains only one class. The GCE will also show the same behaviour in cases where the model over-segments, approaching a full positive mask (all white).

The third group of metrics identified consisted of the precision and the Hausdorff distance. These metrics have been found to be highly sensitive to over-segmentation, leading to strong penalisation of over-segmented masks. This will be the case for the PRC when the positive class is indeed not the majority class, as was the case in our situation and in most applicative contexts: the increased number of FPs will lead to a decrease in the PRC. As for the HDD, which is highly sensitive to distances between clusters of pixels, the distance computed will increase very fast when the detection masks are highly granular and the FP pixels do not exhibit a clear spatial pattern. This metric may be useful in applications where incorrect granular predictions are particularly unappreciated.

The results of the multiple factor analysis on the rankings obtained using three segmentation algorithms (DTSM, SVM, CIVE) demonstrated that the behaviour and the relationships among the metrics are independent of the considered segmentation model and masks it produces, based on its consistent behaviour. If the model consistently over- or under-segments, the rankings of the images produced by the metrics will adapt to the behaviour of the model, given the same set of images, namely given that the metrics roughly adapt to the behaviour of the model, it would be important to identify the metrics or group of metrics that allow answering the following question with the highest possible confidence: "Which model is the best?". As the current work took a "within-model" approach, studying the discriminatory power of the metrics over the set of images, further experimentation needs to be conducted to tackle this "between-models" question. Finally, the current work focused solely on segmentation metrics at the level of pixels. However, segmentation can be studied at multiple levels of abstraction. Further research will be conducted to assess the quality of plant segmentation and detection at the level of "objects" identified in the image, which will allow for a more confident evaluation of quality in cases where the main applicative goal is the detection of these objects, a particularly important task in multiple proximal and remote sensing applications.

Author Contributions: Conceptualisation, P.M., J.-P.D.C., H.E.E. and L.B.; methodology, P.M., J.-P.D.C. and L.B.; software, P.M., E.M. and B.D.; validation, P.M., E.M. and B.D.; formal analysis, P.M., E.M. and B.D.; investigation, P.M., E.M. and B.D.; resources, E.M., B.D. and H.E.E.; data curation, P.M., E.M. and B.D.; writing—original draft preparation, P.M.; writing—review and editing, P.M., J.-P.D.C., E.M., B.D., H.E.E. and L.B.; visualisation, P.M.; supervision, J.-P.D.C., H.E.E. and L.B.; project administration, J.-P.D.C. and H.E.E.; funding acquisition, J.-P.D.C. and H.E.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: For business confidentiality reasons, this paper's data cannot be made public at the moment. They will be made available upon publication of the paper.

Acknowledgments: The authors wish to acknowledge the efforts of the professional agronomists at EXXACT Robotics: Marc Conrad, Guillaume Vecten, and Lucie Dumont, for the acquisition and annotation of the data used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations, in alphabetical order, are used in this manuscript:

ACC	Accuracy
AMI	Adjusted Mutual Information
BAC	Balanced Accuracy
CIVE	Colour Index of Vegetation Extraction
DTSM	Decision Tree Segmentation Model
F1S	F1-Score
GCE	Global Consistency Error
HDD	Hausdorff Distance
IOU	Intersection Over Union
KAP	Cohen's Kappa
MFA	Multiple Factor Analysis
NPV	Negative Predictive Value
PCA	Principal Component Analysis
PRC	Precision
RCL	Recall
RVA	Relative Vegetation Area
SVM	Support Vector Machine

References

1. Salzberg, S.L. On Comparing Classifiers: A Critique of Current Research and Methods. In *Data Mining and Knowledge Discovery*; Kluwer Academic Publishers: Boston, MA, USA, 1999.
2. Zheng, A. *Evaluating Machine Learning Models*; O'Reilly Media, Inc.: Sepastopol, CA, USA, 2015.
3. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence*; Sattar, A., Kang, B.H., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4304, pp. 1015–1021. [CrossRef]
4. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; IEEE: Istanbul, Turkey, 2010; pp. 3121–3124. [CrossRef]
5. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv* **2020**, arXiv:2011.03395.
6. Gudivada, V.; Apon, A.; Ding, J. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *Int. J. Adv. Softw.* **2017**, *10*, 1–20.
7. Breck, E.; Zinkevich, M.; Polyzotis, N.; Whang, S.; Roy, S. Data Validation for Machine Learning. In Proceedings of the SysML, Palo Alto, CA, USA, 31 March–2 April 2019.
8. Jain, A.; Patel, H.; Nagalapatti, L.; Gupta, N.; Mehta, S.; Guttula, S.; Mujumdar, S.; Afzal, S.; Sharma Mittal, R.; Munigala, V. Overview and Importance of Data Quality for Machine Learning Tasks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; KDD '20, pp. 3561–3562. [CrossRef]
9. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
10. Ponti, M.A.; Ribeiro, L.S.F.; Nazare, T.S.; Bui, T.; Collomosse, J. Everything You Wanted to Know about Deep Learning for Computer Vision but Were Afraid to Ask. In Proceedings of the 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Rio de Janeiro, Brazil, 17–20 October 2017; pp. 17–41. [CrossRef]
11. Ouhami, M.; Hafiane, A.; Es-Saady, Y.; El Hajji, M.; Canals, R. Computer Vision, IoT and Data Fusion for Crop Disease Detection Using Machine Learning: A Survey and Ongoing Research. *Remote Sens.* **2021**, *13*, 2486. [CrossRef]
12. Caruana, R.; Niculescu-Mizil, A. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'04, Seattle, WA, USA, 22–25 August 2004; ACM Press: Seattle, WA, USA, 2004; p. 69. [CrossRef]
13. Alaiz-Rodriguez, R.; Japkowicz, N.; Tischer, P. Visualizing Classifier Performance on Different Domains. In Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, OH, USA, 3–5 November 2008; Volume 2, pp. 3–10. [CrossRef]

14. Seliya, N.; Khoshgoftaar, T.M.; Van Hulse, J. A Study on the Relationships of Classifier Performance Metrics. In Proceedings of the 2009 21st IEEE International Conference on Tools with Artificial Intelligence, Newark, NJ, USA, 2–4 November 2009; IEEE: Newark, NJ, USA, 2009; pp. 59–66. [CrossRef]
15. Rakhmatuiln, I.; Kamilaris, A.; Andreasen, C. Deep Neural Networks to Detect Weeds from Crops in Agricultural Environments in Real-Time: A Review. *Remote Sens.* **2021**, *13*, 4486. [CrossRef]
16. Sharma, A.; Jain, A.; Gupta, P.; Chowdary, V. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access* **2021**, *9*, 4843–4873. [CrossRef]
17. Mavridou, E.; Vrochidou, E.; Papakostas, G.A.; Pachidis, T.; Kaburlasos, V.G. Machine Vision Systems in Precision Agriculture for Crop Farming. *J. Imaging* **2019**, *5*, 89. [CrossRef] [PubMed]
18. Barrow, H.G.; Tenenbaum, J.M. Recovering Intrinsic Scene Characteristics from Images. In *Computer Vision Systems*; Academic Press: Waltham, MA, USA, 1978.
19. Fieguth, P. *Statistical Image Processing and Multidimensional Modeling*; Information Science and Statistics; Springer: New York, NY, USA, 2011.
20. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef] [PubMed]
21. Mittal, H.; Pandey, A.C.; Saraswat, M.; Kumar, S.; Pal, R.; Modwel, G. A comprehensive survey of image segmentation: Clustering methods, performance parameters, and benchmark datasets. *Multimed. Tools Appl.* **2021**. [CrossRef] [PubMed]
22. Li, Y.; Huang, Z.; Cao, Z.; Lu, H.; Wang, H.; Zhang, S. Performance Evaluation of Crop Segmentation Algorithms. *IEEE Access* **2020**, *8*, 36210–36225. [CrossRef]
23. Guo, W.; Rage, U.K.; Ninomiya, S. Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model. *Comput. Electron. Agric.* **2013**, *96*, 58–66. [CrossRef]
24. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
25. Breiman, L. (Ed.) *Classification and Regression Trees*, 1st ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 1998.
26. Rico-Fernández, M.; Rios-Cabrera, R.; Castelán, M.; Guerrero-Reyes, H.I.; Juarez-Maldonado, A. A contextualized approach for segmentation of foliage in different crop species. *Comput. Electron. Agric.* **2019**, *156*, 378–386. [CrossRef]
27. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Statistics for Engineering and Information Science; Springer: Berlin, Germany, 2010.
28. Kataoka, T.; Kaneko, T.; Okamoto, H.; Hata, S. Crop growth estimation system using machine vision. In Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003), Kobe, Japan, 20–24 July 2003; Volume 2, pp. b1079–b1083. [CrossRef]
29. Rijsbergen, C.J.V. *Information Retrieval*, 2nd ed.; Butterworth-Heinemann: Newton, MA, USA, 1979.
30. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; IEEE Computer Society: Vancouver, BC, Canada, 2001; Volume 2, pp. 416–423. [CrossRef]
31. Suh, H.K.; Hofstee, J.W.; van Henten, E.J. Improved vegetation segmentation with ground shadow removal using an HDR camera. *Precis. Agric.* **2018**, *19*, 218–237. [CrossRef]
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Vinh, N.X.; Epps, J.; Bailey, J. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; Association for Computing Machinery: New York, NY, USA, 2009; ICML'09, pp. 1073–108. [CrossRef]
34. Hausdorff, F. *Grundzüge der Mengenlehre*; Goschens Lehrbücherei/Gruppe I: Reine und Angewandte Mathematik Series; Von Veit; Verlag von Veit & Comp.: Leipzig, Germany, 1914.
35. Huttenlocher, D.; Klanderman, G.; Rucklidge, W. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 850–863. [CrossRef]
36. De la Torre, F.; Black, M. Robust principal component analysis for computer vision. In Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 1, pp. 362–369. [CrossRef]
37. Heumann, C.; Schomaker, M.; Shalabh. *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*, 2016 ed.; Springer International Publishing: Berlin, Germany, 2016.
38. Chavent, M.; Kuentz-Simonet, V.; Liquet, B.; Saracco, J. ClustOfVar: An R Package for the Clustering of Variables. *J. Stat. Softw.* **2012**, *50*, 1–16. [CrossRef]
39. Joliffe, I. *Principal Component Analysis*; Springer Series in Statistics; Springer: Berlin, Germany, 2002. [CrossRef]
40. Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
41. Lê, S.; Josse, J.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]
42. Escofier, B.; Pagès, J. *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interpétation*; Dunod: Paris, France, 2008.

43. Abdi, H.; Williams, L.J.; Valentin, D. Multiple factor analysis: Principal component analysis for multitable and multiblock datasets: Multiple factor analysis. *WIREs Comput. Stat.* **2013**, *5*, 149–179. [CrossRef]
44. Husson, F.; Josse, J.; Pagès, J. *Principal Component Methods—Hierarchical Clustering—Partitional Clustering: Why Would We Need to Choose for Visualizing Data?* Technical Report; Agrocampus Ouest: Rennes, France, 2021.